Journal of Multivariate Analysis **73**, 120–135 (2000)
doi:10.1006/jmva.1999.1873, available online at http://www.idealibrary.com on IDEAL®

On Parameters of Increasing Dimensions

Xuming He[1]

Department of Statistics, University of Illinois, 725 S. Wright, Champaign, Illinois 61820
E-mail: x-he@uiuc.edu

and

Qi-Man Shao[2]

Department of Mathematics, University of Oregon, Eugene, Oregon 97403
E-mail: shao@math.uoregon.edu

In statistical analyses the complexity of a chosen model is often related to the size of available data. One important question is whether the asymptotic distribution of the parameter estimates normally derived by taking the sample size to infinity for a fixed number of parameters would remain valid if the number of parameters in the model actually increases with the sample size. A number of authors have addressed this question for the linear models. The component-wise asymptotic normality of the parameter estimate remains valid if the dimension of the parameter space grows more slowly than some root of the sample size. In this paper, we consider M-estimators of general parametric models. Our results apply to not only linear regression but also other estimation problems such as multivariate location and generalized linear models. Examples are given to illustrate the applications in different settings. © 2000 Academic Press

AMS 1991 subject classifications: primary 62F12, 62J05; secondary 60F15, 62F35.

Key words and phrases: asymptotic approximation, exponential inequality, increasing dimension, linear regression, logistic regression, M-estimator, self-normalization, spatial median.

1. INTRODUCTION

Complexity of statistical models is often limited by the amount of data we have. When a small parametric model is used, a bias generally exists.

[1] Research supported in part by NSF Grant SBR 9617278.
[2] Research supported in part by NSF Grant DMS 9802451.

0047-259X/00 $35.00

Enlarging the model with more parameters would increase the variability of the estimates. In practice, one tends to choose a larger model when a larger sample size becomes available. This is not usually reflected in the asymptotic analysis in statistics. Typically, a parameter estimate is shown to be approximately normal when the sample size tends to infinity but the dimension of the parameter space is fixed. The asymptotic normality results are highly desirable as they enable us to understand the large sample behavior of the estimates and make large sample inferences about some aspects of the model. For example, we frequently report the standard errors for individual parameter estimates based on their estimated asymptotic variances. It is natural to ask whether they remain valid if the dimension of the parameter space increases with sample size, a situation that is often more relevant to reality.

Huber (1973) considered the asymptotic behavior of M-estimators of linear regression when the number of parameters increases with sample size. Consider the model $y_i = x_i'\beta + e_i$ with $\beta \in R^p$ and e_i as independent errors, and an M-estimator that solves $\sum_i \phi(y_i - x_i'\beta)\, x_i = 0$ for some score function ϕ. Huber (1973) showed that if ϕ is differentiable, the asymptotic normality for $\alpha'\hat{\beta}_n$, where $\alpha \in R^p$, holds if $p^3/n \to 0$ as n increases.

A number of authors have successfully improved on Huber's results. If ϕ is sufficiently smooth, Yohai and Maronna (1979) show that in appropriate balanced cases, $p^{5/2}/n \to 0$ is sufficient for asymptotic normality. Portnoy (1985) and Mammen (1989) weaken the condition to $p^{3/2}\log(p)/n \to 0$. Welsh (1989) considers more general cases where ϕ may have jump discontinuities and shows that $p^3(\log n)^2/n \to 0$ is sufficient. Bai and Wu (1994) further pointed out that the condition on p can be viewed as an integrated part of the design conditions.

Most discussions on the asymptotics of increasing dimensions have centered on the linear regression model. In this paper, we consider M-estimators for more general models with $m = m_n$ dimensional parameters. We extend the results of He and Shao (1996) on Bahadur representation of M-estimators with fixed m to the case of increasing dimensions and show that the asymptotic normality result holds uniformly for any projection of the parameter θ if m grows at a controlled rate. The results we obtain include the linear regression model as a special case, for which our conditions on m or p relative to n are about the same as those of Welsh (1989) whereas our conclusions are slightly stronger. But the main purpose of the present paper is to provide results that apply to other estimation problems. The general results obtained in Section 2 apply to M-estimators of other models including non-linear regression, multivariate location estimation, and generalized linear models. The applicability of our results are illustrated through some examples. Section 3 contains the proofs for the main results.

2. MAIN RESULTS AND APPLICATIONS

We now considers M-estimators for general models. Let $x_1, ..., x_n \in R^p$ be independent observations from probability distribution $F_{i,\theta}$, $i=1, ..., n$, with a common parameter $\theta \in R^m$. Note that m may increase with the sample size n. Sufficient conditions will be found on how fast m can grow without upsetting the asymptotic normality result for an M-estimator of each component of θ or their linear combinations. There are no additional restrictions on the size of p, the dimension of x_i. (But in applications to regression, p and m are of the same magnitude.)

We consider an M-estimator $\hat{\theta}_n$ that minimizes $\sum_{i=1}^n \rho(x_i, \theta)$ over $\theta \in R^m$ for some objective function $\rho(x, \theta)$ that is convex in θ. In this paper, we assume that $\rho(x, \theta) \to +\infty$ as $\|\theta\| \to \infty$ for each x, and ρ is differentiable with respect to θ except at finitely many points. The derivative is denoted by $\psi(x, \theta)$. We define $\|a\|$ to be the L_2 norm of a, and $S_m = \{\alpha \in R^m : \|\alpha\| = 1\}$.

We further assume that the M-estimator satisfies

(C0) $\|\sum_{i=1}^n \psi(x_i, \hat{\theta}_n)\| = o_P(n^{1/2})$.

Remark 2.1. If ρ is differentiable in θ, the left hand side of (C0) is 0. In some special cases of interest where ρ is not everywhere differentiable, (C0) needs to be verified.

Let $\theta_0 \in R^m$ be the solution to $\sum_{i=1}^n E_\theta \psi(x_i, \theta) = 0$, and define

$$\eta_i(\tau, \theta) = \psi(x_i, \tau) - \psi(x_i, \theta) - E\psi(x_i, \tau) + E\psi(x_i, \theta). \tag{2.1}$$

We make additional assumptions as follows, where λ_{min} denotes the smallest eigenvalue of a matrix.

(C1) There exist C and $r \in (0, 2]$ such that $\max_{i \leqslant n} E_\theta \sup_{\tau:\, \|\tau - \theta\| \leqslant d} \|\eta_i(\tau, \theta)\|^2 \leqslant n^C d^r$, for $0 < d \leqslant 1$.

(C2) $\|\sum_{i=1}^n \psi(x_i, \theta_0)\| = O_P((nm)^{1/2})$.

(C3) There exists a sequence of m by m matrices D_n with $\liminf_{n \to \infty} \lambda_{min}(D_n) > 0$ such that for any $B > 0$ and uniformly in $\alpha \in S_m$,

$$\sup_{\|\theta - \theta_0\| \leqslant B(m/n)^{1/2}} \left| \alpha' \sum_{i=1}^n E_{\theta_0}(\psi(x_i, \theta) - \psi(x_i, \theta_0)) - n\alpha' D_n(\theta - \theta_0) \right| = o(n^{1/2}).$$

Condition (C2) is implied by $\sum_{i=1}^n E\|\psi(x_i, \theta_0)\|^2 = O(nm)$. Condition (C3) is often the hardest to check. The key to find how fast m can grow with the sample size for consistency and asymptotic normality is to obtain a tight bound $A(n, m)$ such that the following (C4) and (C5) hold.

(C4) $\sup_{\tau:\, \|\tau-\theta\| \leqslant B(m/n)^{1/2}} \sum_{i=1}^{n} E_\theta\, |\alpha' \eta_i(\tau, \theta)|^2 = O(A(n, m))$ for any $\theta \in R^m$, $\alpha \in S_m$ and $B > 0$.

(C5) $\sup_{\alpha \in S_m} \sup_{\tau:\, \|\tau-\theta\| \leqslant B(m/n)^{1/2}} \sum_{i=1}^{n} (\alpha' \eta_i(\tau, \theta))^2 = O_P(A(n, m))$ for any $\theta \in R^m$ and $B > 0$.

Since we only require $A(n, m)$ to be a bound, it can be obtained in different ways. In most cases, a bound can be obtained easily by expanding $\eta_i(\tau, \theta)$ for each i. For instance, we have

LEMMA 2.1. *If* $E_\theta\{\sup_{\tau:\, \|\tau-\theta\| \leqslant B(m/n)^{1/2}} \|\eta_i(\tau, \theta)\|^2\} \leqslant a_{i,n,m}$, *then* (C4) *and* (C5) *hold with* $A(n, m) = \sum_{i=1}^{n} a_{i,n,m}$.

However, in non-i.i.d. cases tighter bounds on $A(n, m)$ are often possible through the general formulation of (C4) and (C5) with the supremum over τ applied to the summation and with consideration of a projection of ψ_i into any direction α.

Remark 2.2. In typical applications, we can expect $A(n, m) = n(m/n)^{r/2}$, where r measures the smoothness of the score function. See He and Shao (1996) for a class of minimum L_p distance estimators with r ranging from 0 to 1.

Both (C4) and (C5) are easy to check if ψ is Lipschitz in θ. On the other hand, for bounded scores, we have

LEMMA 2.2. *Suppose that* $\sup_{x,\theta} \|\psi(x, \theta)\| \leqslant c_{n,m}$ *for some sequence* $c_{n,m}$, *then* (C5) *is satisfied with the same* $A(n, m)$ *as in* (C4) *provided that* $c_{n,m}^2\, m \log n = O(A(n, m))$.

We now state a rate of convergence result for the estimator.

THEOREM 2.1. *If* $\hat{\theta}_n$ *is a minimizer of* $\sum_i \rho(x_i, \theta)$ *where* ρ *is convex and its derivative with respect to* θ *is* ψ, *then under the assumptions* (C0)–(C5) *with* $A(n, m) = o(n/\log n)$, *we have*

$$\|\hat{\theta}_n - \theta_0\|^2 = O_P(m/n). \tag{2.2}$$

Remark 2.3. As shown in the proof in Section 3, Theorem 2.1 remains true when the error bound $o(n^{1/2})$ of condition (C3) is replaced by $o((nm)^{1/2})$.

Note that the growth rate of m relative to n is governed by the size of $A(n, m)$ we are able to obtain. For example, if $A(n, m) = n(m/n)^{r/2}$, then this requires $m(\log m)^{2/r}/n \to 0$. If we obtain a loose bound for $A(n, m)$, we would need to impose stronger conditions on m.

We now consider the asymptotic distribution of $\alpha'(\hat{\theta}_n - \theta_0)$ for any $\alpha \in S_m$.

THEOREM 2.2. *Suppose that the conditions* (C0)–(C5) *are satisfied with* $A(n, m) = o(n/(m \log n))$. *Then for any consistent estimator* $\hat{\theta}_n$,

$$\hat{\theta}_n - \theta_0 = -n^{-1} \sum_{i=1}^{n} D_n^{-1} \psi(x_i, \theta_0) + r_n,$$

with $\|r_n\| = o_P(n^{-1/2})$.

As a consequence, if x_i are i.i.d variables with $Var(\psi(x_i, \theta_0)) = \Delta$, then $n^{1/2}\alpha'(\hat{\theta}_n - \theta_0)/\sigma_\alpha \to N(0, 1)$ for any $\alpha \in S_m$, where $\sigma_\alpha^2 = \alpha' D_n^{-1} \Delta D_n'^{-1} \alpha$.

We defer the proofs of lemmas and theorems to Section 3. Instead, we now consider three examples to illustrate their applications. The results we get for the second and third examples are new, but we start with the linear regression example which has been studied by other authors.

EXAMPLE 1. Consider the usual linear regression model

$$y_i = z_i'\theta_0 + e_i \tag{2.3}$$

with independent error e_i having a common density f. To use the general setting, identify $x_i = (z_i, y_i)$. The dimension of θ_0 is $m = p - 1$ whose dependence on the sample size n is suppressed in the notation. The design points z_i are either random (and independent of e_i) or fixed. In the latter case one can consider z_i as having point mass. The M-estimator $\hat{\theta}_n$ is a solution to minimizing $\sum_{i=1}^{n} \rho(y_i - z_i'\theta)$ for some convex loss function ρ with minimum at $\rho(0) = 0$. With a possible adjustment of the intercept in the model, we assume without loss of generality that $E\psi(e_i) = 0$. Now consider

$$\left\| \sum_{i=1}^{n} \phi(y_i - z_i'\theta)\, z_i \right\| = O_P(\delta_n) \tag{2.4}$$

with $\phi(x) = \rho'(x)$. If the derivative of ρ exists everywhere, then $\delta_n = 0$. In the case of the least absolute regression with $\rho(x) = |x|$, we have $\delta_n = p^{3/2} \log n$ under appropriate design conditions. For convenience, we consider two types of ϕ functions. If ϕ is Lipschitz and $\delta_n = 0$, we call it a "smooth" score. If $\delta_n = p^{3/2} \log n$ and ϕ has finitely many jump discontinuities but Lipschitz in each interval between two jump points, we call it a "jump" score. To further make the presentation simpler, we state the following assumptions:

(D1) $n^{-1}\sum_i z_i z_i' = I$, where I is the p by p identity matrix.

(D2) ϕ' and ϕ'' are bounded with $c_0 = E\phi'(e_i) \in (0, \infty)$ for smooth scores, or ϕ, f and f' are bounded with $c_0 = -\int_{-\infty}^{\infty} \phi(r) f'(r)\, dr \in (0, \infty)$ for jump scores.

(D3) $\max_{i \leqslant n} \|z_i\|^2 = O(p)$ and $\sup_{\beta, \gamma \in S_m} \sum_{i=1}^n |z_i'\beta|^2 |z_i'\gamma|^2 = O(n)$.

The design condition (D1) may be achieved after a transformation of variables. Condition (D3) is true almost surely if, for example, z_i is a random sample from a p-variate distribution such that $E|\alpha' z|^4$ is uniformly bounded for $\alpha \in S_m$ and for all m. If the distribution of z is spherically symmetric, then it suffices that each component has a finite fourth moment. Also note that the condition (C11) of Welsh (1989) is equivalent to $|\alpha' z_i|$ being uniformly bounded, thus stronger than (D3). We now verify the conditions for Theorem 2.2.

First, for smooth scores where ϕ is Lipschitz such that $|\phi(x) - \phi(y)| \leqslant K|x - y|$ for some $K < \infty$, (C1) holds with $r = 2$ if $p = O(n^{r_1})$ for some $r_1 > 0$, because $\|\eta_i(\tau, \theta)\| \leqslant 2K \|z_i\|^2 \|\tau - \theta\|$ and $\|z_i\|^2 = O(p)$. Similarly due to (D3), (C4) and (C5) hold with $A(n, m) = m$.

For jump scores, it suffices to consider ϕ with only one jump at point 0. Suppose that $|\phi| \leqslant K_1$, $f \leqslant K_1$, and K_2 is a common Lipschitz bound for ϕ on both sides of 0. In this case,

$$\sum_i (\phi(y_i - z_i'\tau) - \phi(y_i - z_i'\theta))^2 (\alpha' z_i)^2$$

$$\leqslant \sum_i (\alpha' z_i)^2 \{4K_1^2 I(|y_i - z_i'\theta| \leqslant |z_i'(\tau - \theta)|) + K_2^2 (z_i'(\tau - \theta))^2\}.$$

Thus, $E_\theta \sum_i \|\alpha' \eta_i(\tau, \theta)\|^2 \leqslant 2 \sum_i (\alpha' z_i)^2 \{4K_1^3 |z_i'(\tau - \theta)| + K_2^2 |z_i'(\tau - \theta)|^2\}$. Together with condition (D3), it implies (C4) with $A(n, m) = n(m/n)^{1/2} = (mn)^{1/2}$. Since ϕ is now bounded and $\|z_i\| = O(p^{1/2})$, it is clear from Lemma 2.2 that (C5) holds with $A(n, m) = (mn)^{1/2} + m^{3/2} \log n$. On the other hand, we have $\sup_{\|\alpha\| = 1} \sum_{i=1}^n (\alpha' \eta_i(\tau, \theta))^4 = O(n)$ in this case, so it follows from Lemma 3.2 that (C5) holds with $A(n, m) = (mn \log n)^{1/2}$. Putting things together, we have verified (C4) and (C5) for $A(n, m) = \min\{(mn)^{1/2} + m^{3/2} \log n, (mn \log n)^{1/2}\}$.

For both types of scores, (C2) holds as long as $E|\phi(e_i)|^2 < \infty$. Moreover, we have

$$\left| \alpha' \sum_{i=1}^n (E(\phi(y_i - z_i'\hat{\theta}_n) - \phi(y_i - z_i'\theta_0))\, z_i) - c_0 n \alpha'(\hat{\theta}_n - \theta_0) \right|$$

$$\leqslant c_1 \sum_{i=1}^n |z_i'\alpha|\, |z_i'(\hat{\theta}_n - \theta_0)|^2$$

for some constant c_1 depending on $\sup_r |\phi''(r)|$ for smooth scores or $\sup_r |f'(r)|$ for jump scores. It is then immediate that (C3) holds true. Therefore, we have

COROLLARY 2.1. *Assume that conditions* (D1)–(D3) *are satisfied. We have* $\|\hat{\theta}_n - \theta_0\|^2 = O_P(p/n)$ *for both types of scores provided that* $p(\log p)^3 = o(n)$. *Furthermore, if* $p^2(\log p) = o(n)$ *for smooth scores or* $p^3(\log p)^2 = o(n)$ *for jump scores, then*

$$n^{1/2}\alpha'(\hat{\theta}_n - \theta_0)/\sigma(\alpha) \to N(0, 1), \tag{2.5}$$

for any $\alpha' = (\alpha_1, ..., \alpha_p) \in R^p$, *where*

$$\sigma^2(\alpha) = (c_0 E\phi^2(e))^{-1} \|\alpha\|^2.$$

Under essentially the same conditions on p and slightly stronger conditions on the design, Welsh (1989, 1990) established the asymptotic normality (2.5) for any α such that $\|\alpha\|_1 = \sum_{i=1}^p |\alpha_i| = O(1)$ as $p \to \infty$. So the result we prove here is slightly more general in this aspect. Without imposing other restrictions on the error distribution and on the score function, the condition on p as stated above is the best available in the literature so far. On the other hand, Portnoy (1985) showed that $p^{3/2} \log p/n \to 0$ is sufficient if ϕ is three times differentiable and the error distribution is symmetric.

If, in lieu of Condition (D1), we have $n^{-1} \sum_i z_i z_i' = D_n$, where $0 < \liminf_n \lambda_{min}(D_n) \leqslant \limsup_n \lambda_{max}(D_n) < \infty$, (2.5) holds with $\sigma^2(\alpha) = (c_0 E\phi^2(e))^{-1} \alpha' D_n^{-1} \alpha$. We shall point out that the design conditions we use in the present paper are *not* meant to be the weakest possible. They are chosen partly for simplicity and ease of interpretation. For some results under more general design conditions, see Bai and Wu (1994).

EXAMPLE 2. Spatial median for multivariate data.

Now suppose that a random sample $x_1, x_2, ..., x_n$ of dimension m is used to estimate the multivariate location parameter by minimizing $\sum_{i=1}^n \|x_i - \theta\|$. Suppose that the underlying distribution for x_i has a continuous density respect to Lebesgue measure, then $E\|X - \theta\|$ is convex in θ with its minimum reached at say θ_0.

The spatial median is an M-estimator with

$$\psi(x, \theta) = -(x - \theta)/\|x - \theta\|.$$

We assume

(E1) $E_{\theta_0}(1/\|x - \theta\|^2) = O(1)$ as $m \to \infty$ if $\|\theta - \theta_0\| \leqslant c_1$ for some $c_1 > 0$.

(E2) $\liminf_{m \to \infty} \inf_{\|\alpha\| = 1} E_{\theta_0}\{\|x - \theta_0\|^{-1} - (\alpha'(x - \theta_0))^2/\|x - \theta_0\|^3\} > 0$.

By straightforward calculations,

$$\psi(x,\tau)-\psi(x,\theta)=\frac{1}{\|x-\theta\|}\left(\tau-\theta+\frac{x-\tau}{\|x-\tau\|}\left(\|x-\tau\|-\|x-\theta\|\right)\right),$$

which implies $\sup_{\|\tau-\theta\|\leqslant d}\|\psi(x,\tau)-\psi(x,\theta)\|\leqslant 2d/\|x-\theta\|$. Therefore (C1) holds with $r=2$ and (C4)–(C5) hold with $A(n,m)=m$. Condition (C2) holds automatically.

To verify (C3), we first consider the derivatives of $\alpha'\psi$ with respect to θ. The first order derivative is $L_1(\theta)=\|x-\theta\|^{-1}\alpha-\alpha'(x-\theta)(x-\theta)\|x-\theta\|^{-3}$, and the second order derivative matrix is $L_2(\theta)=\|x-\theta\|^{-3}\alpha'(x-\theta)\times\{I-3\|x-\theta\|^{-2}(x-\theta)(x-\theta)'\}$. It is then clear that $\sup_{\beta\in S_m}\beta' E_{\theta_0}L_2(\theta)\beta=O(1)$ for $\|\theta-\theta_0\|\leqslant c_1$. Then, we have

$$\alpha' E(\psi(x,\theta)-\psi(x,\theta_0))=\alpha' D(\theta-\theta_0)+O(\|\theta-\theta_0\|^2)$$

where $D=E\{(\|x-\theta_0\|^2 I-(x-\theta_0)(x-\theta_0)')/\|x-\theta_0\|^3\}$. So (C3) holds when $m^2/n\to 0$.

COROLLARY 2.2. *Let $\theta_0\in R^m$ be the unique minimizing point of $E\|x-\theta_0\|$. If $m(\log m)^2/n\to 0$, then the spatial median satisfies $\|\hat\theta_n-\theta_0\|^2=O_P(m/n)$. Furthermore, assume* (E1), (E2) *and $m^2\log m=o(n)$. Then*

$$n^{1/2}(\hat\theta_n-\theta_0)=-n^{-1/2}D^{-1}\sum_{i=1}^{n}\frac{x_i-\theta_0}{\|x_i-\theta_0\|}+r_n$$

with $\|r_n\|=o_P(1)$.

It is clear that the requirement we need for the dimension m is the same as that for the smooth M-estimators in regression. The conditions (E1) and (E2) are satisfied for a large class of multivariate distributions.

EXAMPLE 3. Logistic regression. Suppose that the response variable y is now binary so that $P(y=1\,|\,z)=e^{z'\beta_0}/(1+e^{z'\beta_0})$ and $P(y=0\,|\,z)=1/(1+e^{z'\beta_0})$, where z is covariate of dimension p. Given a random sample (z_1,y_1), ..., (z_n,y_n) the log likelihood is

$$\log L(\beta)=\sum_{i=1}^{n}(y_i(z_i'\beta)-\log(1+e^{z_i'\beta}))$$

which is convex in β. The score function is $\psi(x,\beta)=(y-e^{z'\beta}/(1+e^{z'\beta}))\,z$. We assume the following design conditions.

(F1) $\sum_{i=1}^n \|z_i\|^2 = O(np)$, and $\sup_{\alpha, \beta \in S_p} \sum_{i=1}^n |\alpha' z_i|^2 \, |\beta' z_i|^2 = O(n)$.

(F2) $\liminf_{n\to\infty} \lambda_{min}(D_n) > 0$, where $D_n = n^{-1} \sum_{i=1}^n (e^{z_i' \beta_0} / (1 + e^{z_i' \beta_0}))\, z_i z_i'$.

The derivative of ψ with respect to β is $-e^{z'\beta}(1+e^{z'\beta})^{-2}\, zz'$. By assumption (F1), conditions (C1), (C4) and (C5) hold with $r=2$ and $A(n, m)=m$. Condition (C2) is also due to (F1). To verify (C3), note that the 2nd derivative of ψ is $b(z'\beta)\, zz'$ where b is a function bounded by 1. Expanding to the quadratic term, we obtain

$$\left|\sum_i \alpha' E\{\psi(x_i, \beta) - \psi(x_i, \beta_0)\} - n\alpha' D_n(\beta - \beta_0)\right|$$

$$\leqslant (\beta - \beta_0)' \sum_{i=1}^n |\alpha' z_i| \, z_i z_i' (\beta - \beta_0).$$

Assumptions (F1) and (F2) imply (C3). As a result, we have $\|\hat{\beta}_n - \beta_0\|^2 = O_P(p/n)$ if $p \log p/n \to 0$ and the asymptotic normality of $\alpha'(\hat{\beta}_n - \beta_0)$ in logistic regression holds provided that $p^2(\log p)/n \to 0$.

Given the design z_i, the logistic model falls into the family of exponential distributions. Portnoy (1988) studied the dimension asymptotics for exponential families, but assumed i.i.d observations so that it does not cover this example directly.

3. PROOFS OF MAIN RESULTS

Proof of Lemma 2.1. The proof is straightforward by

$$\sup_{\alpha \in S_m} \sup_{\tau:\, \|\tau - \theta\| \leqslant B(m/n)^{1/2}} \sum_{i=1}^n (\alpha' \eta_i(\tau, \theta))^2 \leqslant \sum_{i=1}^n \sup_{\tau:\, \|\tau - \theta\| \leqslant B(m/n)^{1/2}} \|\eta_i(\tau, \theta)\|^2.$$

and by Chebyshev inequality. ∎

Proof of Lemma 2.2. Let $A_n = \{\tau \in R^m : \|\tau - \theta\| \leqslant B((m/n)^{1/2}\}$. It follows from Lemma 3.2 that

$$\sup_{\tau \in A_n} \frac{|\sum_{i=1}^n \{(\alpha' \eta_i(\tau, \theta))^2 - E(\alpha' \eta_i(\tau, \theta))^2\}|}{n^{-2} + (\sum_{i=1}^n E(\alpha' \eta_i(\tau, \theta))^4)^{1/2} + (\sum_{i=1}^n (\alpha' \eta_i(\tau, \theta))^4)^{1/2}}$$

$$= O_P((m \log n)^{1/2}).$$

Thus, by the assumption that $\|\eta_i\| \leqslant c_{n,m}$,

$$\sup_{\tau \in A_n} \frac{|\sum_{i=1}^n \{(\alpha'\eta_i(\tau,\theta))^2 - E(\alpha'\eta_i(\tau,\theta))^2\}|}{n^{-2} + c_{n,m}(\sum_{i=1}^n E(\alpha'\eta_i(\tau,\theta))^2)^{1/2} + c_{n,m}(\sum_{i=1}^n (\alpha'\eta_i(\tau,\theta))^2)^{1/2}}$$
$$= O_P((m \log n)^{1/2}).$$

It is easy to see that for nonnegative numbers x, a and b,

$$\frac{|x-a|}{b+x^{1/2}} \leqslant c \qquad \text{implies} \quad x \leqslant 2a^2 + 4c^2 + 2b^2.$$

Hence it follows from the above that

$$\sup_{\tau \in A_n} \sum_{i=1}^n (\alpha'\eta_i(\tau,\theta))^2 = O_P(A(n,m) + c_{n,m}^2 m \log n).$$

All the O_P bounds above are uniform in $\alpha \in R_m$. ∎

The proofs of Theorems 2.1 and 2.2 are based on Lemma 3.3 below. It has some similarity with Lemma 4.1 of He and Shao (1996), but there are some essential differences. One main difficulty comes from allowing the parameter θ and the score function ψ to vary with n. The basic technique used in He and Shao (1996) is to divide the parameter space into smaller cubes. The total number of cubes grows at a polynomial rate so that the exponential bound obtained at each cube holds globally. However, a straightforward modification of He and Shao (1996) does not lead to the desired result due to the exponentially increasing size of the sub-cubes needed in the chaining argument. The proof in the present paper relies on the following Lemma 3.1 which is of independent interest itself. One can view Lemma 3.1 as an alternative to the well-known Kolmogorov exponential inequality. The latter requires boundness of the variables. We do away with the boundness assumption by considering self-normalization. We prove Lemmas 3.1 and 3.2 first. The proofs of Theorems 2.1 and 2.2 are sketched at the end.

LEMMA 3.1. *Let $X_1, X_2, ..., X_n$ be independent R^d-valued random variables with $EX_i = 0$ and $E\|X_i\|^2 < \infty$ for every $i = 1, 2, ..., n$. Then there exists an absolute constant A such that for any $a > 0$*

$$P\left(\max_{1 \leqslant k \leqslant n} \left\|\sum_{i=1}^k X_i\right\| \geqslant a\left(B_n + \left(\sum_{i=1}^n \|X_i\|^2\right)^{1/2}\right)\right) \leqslant A \exp(-a^2/A), \qquad (3.1)$$

where $B_n = (\sum_{i=1}^n E\|X_i\|^2)^{1/2}$.

Proof. Let $A=300$. When $a\leqslant 15$, then $A\exp(-a^2/A)>1$ and hence (3.1) is trivial. We only need to consider the case of $a>15$. Let $\{Y_i, 1\leqslant i\leqslant n\}$ be an independent copy of $\{X_i, 1\leqslant i\leqslant n\}$. Then, by Chebyshev's and Etemadi's inequalities (see Billingsley (1995), p. 288),

$$\begin{aligned}
&P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} Y_i\right\|\leqslant 12B_n, \sum_{i=1}^{n}\|Y_i\|^2\leqslant 2B_n^2\right)\\
&\quad\geqslant 1-P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} Y_i\right\|>12B_n\right)-P\left(\sum_{i=1}^{n}\|Y_i\|^2>2B_n^2\right)\\
&\quad\geqslant 1-3\max_{1\leqslant k\leqslant n} P\left(\left\|\sum_{i=1}^{k} Y_i\right\|>4B_n\right)-1/2\\
&\quad\geqslant 1-3/16-(1/2)>1/4.
\end{aligned}\tag{3.2}$$

Let $\{\varepsilon_i, 1\leqslant i\leqslant n\}$ be a Rademacher sequence independent of $\{X_i, 1\leqslant i\leqslant n\}$ and $\{Y_i, 1\leqslant i\leqslant n\}$. Also for convenience, let $Q_n=(a/2)(\sum_{i=1}^{n}\|X_i-Y_i\|^2)^{1/2}$. Noting that

$$\begin{aligned}
&\left\{\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} X_i\right\|\geqslant a\left(B_n+\left(\sum_{i=1}^{n}\|X_i\|^2\right)^{1/2}\right),\right.\\
&\qquad\left.\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} Y_i\right\|\leqslant 12B_n, \sum_{i=1}^{n}\|Y_i\|^2\leqslant 2B_n^2\right\}\\
&\quad\subset\left\{\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k}(X_i-Y_i)\right\|\geqslant Q_n\right\},
\end{aligned}$$

and that $\{X_i-Y_i, 1\leqslant i\leqslant n\}$ is a sequence of independent symmetric random variables, we have

$$\begin{aligned}
&P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} X_i\right\|\geqslant a\left(B_n+\left(\sum_{i=1}^{n}\|X_i\|^2\right)^{1/2}\right)\right)\\
&\quad=P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} X_i\right\|\geqslant a\left(B_n+\left(\sum_{i=1}^{n}\|X_i\|^2\right)^{1/2}\right),\right.\\
&\qquad\left.\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} Y_i\right\|\leqslant 12B_n, \sum_{i=1}^{n}\|Y_i\|^2\leqslant 2B_n^2\right)\\
&\qquad\times\left\{P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k} Y_i\right\|\leqslant 12B_n, \sum_{i=1}^{n}\|Y_i\|^2\leqslant 2B_n^2\right)\right\}^{-1}\\
&\quad\leqslant 4P\left\{\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k}(X_i-Y_i)\right\|\geqslant Q_n\right\}
\end{aligned}$$

$$
\begin{aligned}
&= 4P\left\{\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k}(X_i-Y_i)\,\varepsilon_i\right\|\geqslant Q_n\right\}\\
&= 4E\left\{P\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{i=1}^{k}(X_i-Y_i)\,\varepsilon_i\right\|\geqslant Q_n \mid X_i,\, Y_i,\, 1\leqslant i\leqslant n\right)\right\}\\
&\leqslant 12E\left\{\max_{1\leqslant k\leqslant n} P\left(\left\|\sum_{i=1}^{k}(X_i-Y_i)\,\varepsilon_i\right\|\geqslant Q_n/3 \mid X_i,\, Y_i,\, 1\leqslant i\leqslant n\right)\right\}\\
&\leqslant 24\exp\left(-\frac{(a/6)^2}{32}\right),
\end{aligned}
$$

where the last inequality follows from the exponential inequality for the Rademacher sequence (see, e.g., Ledoux and Talagrand (1991), p. 101). This proves (3.1). ∎

LEMMA 3.2. *Let* $\{v_i(t), t\in R^m\}$, $1\leqslant i\leqslant n$ *be independent* R^p*-valued random variables with* $Ev_i(t)=0$ *for all* t. *Assume that there exist* $r_1>0$ *and* $r_2>0$ *such that for every* $s\in R^m$, $0<d\leqslant 1$, $1\leqslant i\leqslant n$

$$
E\sup_{t:\,\|t-s\|\leqslant d}\|v_i(t)-v_i(s)\|\leqslant n^{r_1}d^{r_2}. \tag{3.2}
$$

Let

$$
B_n(t,s)=\left(\sum_{i=1}^{n}E\,\|v_i(t)-v_i(s)\|^2\right)^{1/2} \quad \textit{and}
$$

$$
V_n(t,s)=\left(\sum_{i=1}^{n}\|v_i(t)-v_i(s)\|^2\right)^{1/2}.
$$

Then

$$
\sup_{\|t\|\leqslant n^{r_3},\,\|s\|\leqslant n^{r_3}}\frac{\|\sum_{i=1}^{n}(v_i(t)-v_i(s))\|}{n^{-2}+B_n(t,s)+V_n(t,s)}=O_P((m\log(n+m))^{1/2}) \tag{3.3}
$$

for every $r_3\geqslant 0$.

Proof. It suffices to show that

$$
P(\mathcal{O}_{n,a})=o(1)\qquad \text{as}\quad n\to\infty, \tag{3.4}
$$

for sufficiently large a, where

$$
\mathcal{O}_{n,a}
$$

$$
=\left\{\sup_{t,s\in R^m:\,\|t\|\leqslant n^{r_3},\,\|s\|\leqslant n^{r_3}}\frac{\|\sum_{i=1}^{n}(v_i(t)-v_i(s))\|}{(n^{-2}+B_n(t,s)+V_n(t,s))(m\log(n+m))^{1/2}}>2a\right\}.
$$

Let $Z_n(t, s) = \|\sum_{i=1}^n (v_i(t) - v_i(s))\|$ and

$$\delta := \delta_n = n^{-\varepsilon_0}/m, \qquad M := n^{r_3+1}/\delta = n^{1+r_3+\varepsilon_0}\, m, \tag{3.5}$$

where $\varepsilon_0 = (9 + r_1 + r_2)/r_2$. Consider the concentric cubes

$$\mathscr{C}_l = \{t\colon |t| \leqslant l\delta\}, \qquad l = 1, 2, ..., M,$$

where $|t|$ denotes the maximum norm of t. Subdivide the difference $\mathscr{C}_{l+1} \backslash \mathscr{C}_l$ into smaller cubes with edges of the length δ. For each value of l there are $m_l = (2(l+1))^m - (2l)^m$ such small cubes which are denoted by $\mathscr{C}_l^j$, $j = 1, 2, ..., m_l$. Let c_l^j be the center of $\mathscr{C}_l^j$. Set $\mathscr{C}_0 = \varnothing$. Then for $x > 1$,

$$\left\{ \sup_{\|t\| \leqslant n^{r_3},\ \|s\| \leqslant n^{r_3}} \frac{Z_n(t, s)}{n^{-2} + B_n(t, s) + V_n(t, s)} > 2x \right\}$$

$$\subset \bigcup_{0 \leqslant l < M} \bigcup_{j \leqslant m_l} \bigcup_{0 \leqslant k < M} \bigcup_{i \leqslant m_k} \left\{ \sup_{t \in \mathscr{C}_l^j,\, s \in \mathscr{C}_k^i} \frac{Z_n(t, s)}{n^{-2} + B_n(t, s) + V_n(t, s)} \geqslant 2x \right\}.$$

Let

$$\mathscr{B}_u(l, j; k, i) = \left\{ \sup_{t \in \mathscr{C}_l^j} \|v_u(t) - v_u(c_l^j)\| \leqslant n^{-4},\ \sup_{s \in \mathscr{C}_k^i} \|v_u(s) - v_u(c_k^i)\| \leqslant n^{-4} \right\},$$

$$v_u(l, j; k, i) = (v_u(c_l^j) - v_u(c_k^i))\, I\{\mathscr{B}_u(l, j; k, i)\},$$

$$v_u^*(l, j; k, i) = v_u(l, j; k, i) - E v_u(l, j; k, i).$$

By the Hölder inequality and (3.2), we have

$$\begin{aligned}
&\|E(v_u(t) - v_u(s))\, I\{\mathscr{B}_u^c(l, j; k, i)\}\| \\
&\quad = \|E(v_u(t) - v_u(s))\, I\{\mathscr{B}_u^c(l, j; k, i)\}\| \\
&\quad \leqslant E\, \|v_u(t) - v_u(s)\|\, I\{\mathscr{B}_u^c(l, j; k, i)\} \\
&\quad \leqslant (E\, \|v_u(t) - v_u(s)\|^2)^{1/2}\, (P(\mathscr{B}_u^c(l, j; k, i)))^{1/2} \\
&\quad \leqslant (E\, \|v_u(t) - v_u(s)\|^2)^{1/2}\, (n^{3+r_1}(m\delta)^{r_2})^{1/2} \\
&\quad \leqslant n^{-3} (E\, \|v_u(t) - v_u(s)\|^2)^{1/2}.
\end{aligned}$$

For $t \in \mathscr{C}_l^j$, $s \in \mathscr{C}_k^i$, observe that

$$\begin{aligned}
&\frac{Z_n(t, s)}{n^{-2} + B_n(t, s) + V_n(t, s)} \\
&\quad \leqslant \frac{\left(\begin{array}{l} \|\sum_{u=1}^n (v_u(t) - v_u(s))\, I\{\mathscr{B}_u(l, j; k, i)\}\| \\ \qquad + \|\sum_{u=1}^n (v_u(t) - v_u(s))\, I\{\mathscr{B}_u^c(l, j; k, i)\}\| \end{array}\right)}{n^{-2} + B_n(t, s) + V_n(t, s)} \\
&\quad \leqslant \frac{\|\sum_{u=1}^n (v_u(t) - v_u(s))\, I\{\mathscr{B}_u(l, j; k, i)\}\|}{n^{-2} + B_n(t, s) + V_n(t, s)} + \left(\sum_{u=1}^n I\{\mathscr{B}_u^c(l, j; k, i)\}\right)^{1/2} \\
&\quad \leqslant \frac{\left(\begin{array}{l} \|\sum_{u=1}^n ((v_u(t) - v_u(s)) - E(v_u(t) - v_u(s))) \\ \qquad \times I\{\mathscr{B}_u(l, j; k, i)\}\| + n^{-2} B_n(t, s) \end{array}\right)}{n^{-2} + B_n(t, s) + V_n(t, s)} \\
&\qquad + \left(\sum_{u=1}^n I\{\mathscr{B}_u^c(l, j; k, i)\}\right)^{1/2} \\
&\quad \leqslant \frac{2n^{-3} + \|\sum_{u=1}^n v_u^*(l, j; k, i)\|}{\left(\begin{array}{l} n^{-2} - 4n^{-3} + (1/2)(\sum_{u=1}^n E\, \|v_u^*(l, j; k, i)\|^2)^{1/2} \\ \qquad + (\sum_{u=1}^n \|v_u^*(l, j; k, i)\|^2)^{1/2} \end{array}\right)} \\
&\qquad + n^{-2} + \left(\sum_{u=1}^n I\{\mathscr{B}_u(l, j; k, i)\}\right)^{1/2}. \qquad (3.6)
\end{aligned}$$

Hence for $n \geqslant 16$

$$\begin{aligned}
P(\mathscr{O}_{n,a}) \leqslant{} & \sum_{0 \leqslant l < M} \sum_{j \leqslant m_l} \sum_{0 \leqslant k < M} \sum_{i \leqslant m_i} P\left\{\frac{\|\sum_{u=1}^n v_u^*(l, j; k, i)\|}{\left(\begin{array}{l} (\sum_{u=1}^n E\, \|v_u^*(l, j; k, i)\|^2)^{1/2} \\ \qquad + (\sum_{u=1}^n \|v_u^*(l, j; k, i)\|^2)^{1/2} \end{array}\right)}\right. \\
& \qquad \left. \geqslant a(m \log(n+m))^{1/2}/8 \right\} \\
& + \sum_{0 \leqslant l < M} \sum_{j \leqslant m_l} \sum_{0 \leqslant k < M} \sum_{i \leqslant m_i} P\left\{\sum_{u=1}^n I\{\mathscr{B}_u(l, j; k, i)\} \geqslant a^2 m \log(n+m)\right\} \\
\leqslant{} & A \sum_{0 \leqslant l < M} \sum_{j \leqslant m_l} \sum_{0 \leqslant k < M} \sum_{i \leqslant m_i} \exp(-A^{-1} a^2 m \log(n+m)) \\
& \qquad [\text{by Lemma 3.1}] \\
& + \sum_{0 \leqslant l < M} \sum_{j \leqslant m_l} \sum_{0 \leqslant k < M} \sum_{i \leqslant m_i} \left(\frac{3 \sum_{u=1}^n P(\mathscr{B}_u^c(l, j; k, i))}{a^2 m \log(n+m)}\right)^{a^2 m \log(n+m)}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant A \exp(-m \log(n+m)) \\
&\quad + \sum_{0 \leqslant l < M} \sum_{j \leqslant m_l} \sum_{0 \leqslant k < M} \sum_{i \leqslant m_i} \left(\frac{3 \sum_{u=1}^{n} n^{4+r_1} (m\delta)^{r_2}}{a^2 m \log(n+m)} \right)^{a^2 m \log(n+m)} \\
&\leqslant 2A \exp(-m \log(n+m)), \qquad (3.7)
\end{aligned}
$$

provided that a is sufficiently large, where the third last inequality comes from the exponential inequality for the binomial distribution (see Ledoux and Talagrand (1991), p. 51). This proves (3.7). ∎

An immediate consequence of Lemma 3.2 is

LEMMA 3.3. *Let* $R_n = (m \log(n+m))^{1/2}$ *Under the condition* (C1), *we have*

$$
\sup_{\|\tau\|,\, \|\theta\| \leqslant n^{r_3}} \frac{\|\sum_{i=1}^{n} \eta_i(\tau, \theta)\|}{n^{-2} + (\sum_{i=1}^{n} E\, \|\eta_i(\tau, \theta)\|^2)^{1/2} + (\sum_{i=1}^{n} \|\eta_i(\tau, \theta)\|^2)^{1/2}} = O_P(R_n), \tag{3.8}
$$

$$
\sup_{\|\tau\|, \|\theta\| \leqslant n^{r_3}} \frac{|\sum_{i=1}^{n} (\tau - \theta)' \, \eta_i(\tau, \theta)|}{\begin{pmatrix} n^{-2} + (\sum_{i=1}^{n} E\, |(\tau - \theta)' \, \eta_i(\tau, \theta)|^2)^{1/2} \\ + (\sum_{i=1}^{n} |(\tau - \theta)' \, \eta_i(\tau, \theta)|^2)^{1/2} \end{pmatrix}} = O_P(R_n), \tag{3.9}
$$

and

$$
\sup_{\|\alpha\|,\, \|\tau\|,\, \|\theta\| \leqslant n^{r_3}} \frac{|\sum_{i=1}^{n} \alpha' \eta_i(\tau, \theta)|}{\begin{pmatrix} n^{-2} + (\sum_{i=1}^{n} E\, |\alpha' \eta_i(\tau, \theta)|^2)^{1/2} \\ + (\sum_{i=1}^{n} |\alpha' \eta_i(\tau, \theta)|^2)^{1/2} \end{pmatrix}} = O_P(R_n) \tag{3.10}
$$

for every $r_3 > 0$, *where* $\eta_i(\tau, \theta)$ *is as defined in* (2.1).

Proof of Theorem 2.1. Due to the convexity of the objective function, it suffices to show that for any $\varepsilon > 0$, there exists $B < \infty$ such that $P\{\inf_{\|\eta\|=1} \sum_i \eta' \psi(x_i, \theta_0 + B(m/n)^{1/2}) \, \eta) > 0\} > 1 - \varepsilon$ for sufficiently large n.

By (3.9) and Conditions (C3), (C4) and (C5), $\sum_i \eta' \psi(x_i, \theta_0 + B(m/n)^{1/2} \eta) = \sum_i \eta' \psi(x_i, \theta_0) + B(mn)^{1/2} \eta' D_n \eta + o(n^{1/2}) + O_P((A(n, m)\, m \log n)^{1/2})$ uniformly in $\eta \in S_m$. If $A(n, m) = o(n/\log n)$, the remainder term is $o((mn)^{1/2})$. Thus for sufficiently large n,

$$
\begin{aligned}
&\left\{ \inf_{\|\eta\|=1} \sum_i \eta' \psi(x_i, \theta_0 + B(m/n)^{1/2} \, \eta) > 0 \right\} \\
&\qquad \supset \left\{ (mn)^{-1/2} \inf_{\|\eta\|=1} \sum_i \eta' \psi(x_i, \theta_0) > -(B/2) \, \lambda_{min}(D_n) \right\}
\end{aligned}
$$

whose probability tends to 1 due to (C2) as $B, n \to \infty$. ∎

Proof of Theorem 2.2. We only need to consider α such that $\|\alpha\| = O(1)$ as $m \to \infty$. By (3.10) and conditions (C3), (C4), and (C5), we have

$$\sum_i \alpha' \psi(x_i, \theta_0) + n\alpha' D_n(\hat{\theta}_n - \theta_0)$$
$$= o_P(n^{1/2}) + O_P((A(n, m)\, m \log n)^{1/2}) = o_P(n^{1/2})$$

if $A(n, m) = o(n/(m \log n))$. This representation is uniform in α as long as $\|\alpha\|$ is bounded. Now $\liminf_n \lambda_{min}(D_n) > 0$, we can replace α by $D_n^{-1}\alpha$ to get the desired result in Theorem 2.2. ▮

ACKNOWLEDGMENT

We wish to thank two referees and an Editor for their helpful comments and suggestions on our earlier version of the paper.

REFERENCES

1. Z. Bai and Y. Wu, Limiting behavior of M-estimators of regression coefficients in high dimensional linear models, I. Scale-dependent case, *J. Multivariate Anal.* **51** (1994), 211–239.
2. P. Billingsley, "Probability and Measure," Wiley, New York, 1995.
3. X. He and Q. M. Shao, A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs, *Ann. Statist.* **24** (1996), 2608–2630.
4. P. J. Huber, Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.* **1** (1973), 799–821.
5. Ledoux and Talagrand, "Probability in Banach Spaces," Springer-Verlag, New York, 1991.
6. E. Mammen, Asymptotics with increasing dimension for robust regression with applications to the bootstrap, *Ann. Statist.* **17** (1989), 382–400.
7. S. Portnoy, Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: II. Normal approximation, *Ann. Statist.* **13** (1985), 1403–1417.
8. S. Portnoy, Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Ann. Statist.* **16** (1988), 356–366.
9. A. H. Welsh, On M-processes and M-estimation, *Ann. Statist.* **17** (1989), 337–361.
10. A. H. Welsh, Correction: "On M-processes and M-estimation," *Ann. Statist.* **18** (1990), 1500.